HW4 - COMP 329: NLP

Matt Hyatt

April 10, 2022

Procedure

1

I trained my word2vec model on the OANC corpus from http://www.anc.org/

I validated the model with the following examples:

Word One	Word Two	Similarity
apple	orange	0.52
apple	sauce	0.54
apple	pie	0.58
apple	tree	0.39
apple	banana	0.45
apple	orchard	0.38
apple	running	0.12
cake	pie	0.56
teacher	student	0.73

2

I tested the pretrained model from the Google News corpus to find groups of similar words. The top 5 most similar words for each group were:

- car:
  - truck, cars, garage, vehicle, boat
  - all words relating to vehicles
- fruit:
  - vegetables, meat, vegetable, flowers, cheese
  - different kinds of foods (except flowers, when generate fruit when pollenated)
- exercise:
  - schedule, improvement, aerobic, environment, ventilation
  - people often schedule workouts and use them for self improvement. Cardiovascular exercise is aerobic. I'm not sure why environment and ventilation were included.
- president:
  - presidency, versa, administration, jiang, hillary
  - Jiang Zemin and Hillary Clinton are both political figures (presidents or presidential candidates). Presidents serve as administration for a country. I do not know why versa was included.

- earthquake:
  - explosion, outbreak, disaster, eruption, quake
  - All are examples of natural disasters or destructive forces.

3

My pretrained model recieved a spearman correlation coefficient of 0.57 on the WordSim-353 gold standard and a p value of 7.58e-19.

4

I proposed the following word analogies and tested them using `wv.similarity()`:

Analogy	Similarity
swimmer - swim + run = runner	0.29
baker - bread + money = banker	0.18
fruit - sweet + sour = lemon	0.40
moon - night + day = sun	0.21

I was surprised how similar the third example was given that fruit - sweet + sour seems to be quite general and lemons are already fruits. I assumed that the swimmer and the moon analogies would recieve higher cosine similarity since the logic of the analogy seems more coherent to me.